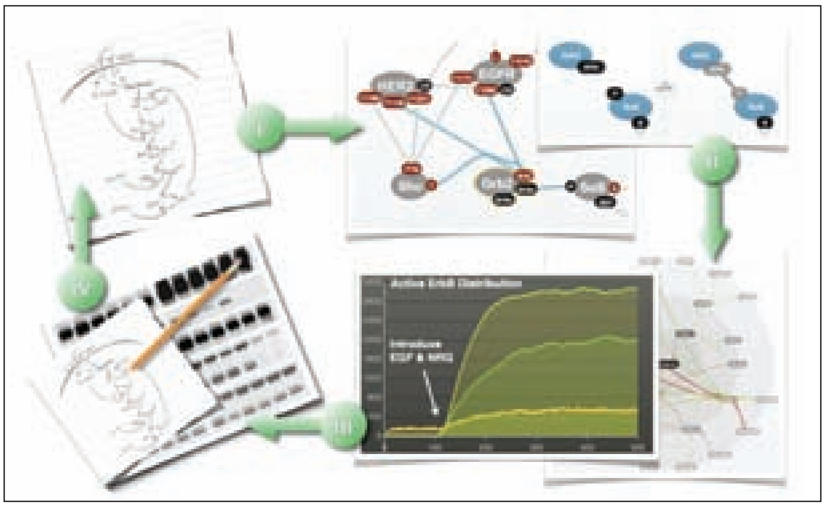
BIOLOGISTS FLIRT WITH MODELS

The enormous challenge posed by the complexity of biological systems represents a potential intellectual impasse to researchers and threatens to stall future progress in basic biology and healthcare. In recent years, increasing reliance on correlative approaches to biology has failed to resolve this situation. The burgeoning volumes of laboratory data gathered in support of these approaches pose more questions than they answer until such time as they can be assimilated as real knowledge. Modelling can provide the kind of intellectual frameworks needed to transform data into knowledge, yet very few modelling methodologies currently exist that are applicable to the large, complex systems of interest to biologists. Early developments in biological modelling, driven largely by the convergence in systems biology of complementary approaches from other disciplines, hold the potential to forge a knowledge revolution in biology and to bring modelling into the mainstream of biological research in the way that it is in other fields. This will require the development of modelling platforms that allow biological systems to be described using idioms familiar to biologists, that encourage experimentation and that leverage the connectivity of the internet for scientific collaboration and communication.

By Gordon Webster

Researchers in biology and the life sciences have yet to embrace modelling in the way that their peers in the fields of physics, chemistry and engineering have done. There are some very good reasons for this, not the least of which is that the systems of interest to biologists tend to be far more refractory to modelling than the systems that are studied in these other fields. It is much more straightforward, for example, to capture in lines of computer code the orbit of a satellite or the load on a suspension bridge than it is to build a computational model of even the simplest and most well-studied of biological organisms. At the molecular level the fundamental processes that occur in living systems can also be described in terms of physics and chemistry. However, at the more macroscopic scales at which these systems can be studied as 'biological' entities, the researcher is confronted with an enormous number of moving parts, a web of interactions of astronomical complexity, significant heterogeneity between the many 'copies' of the system and a degree of stochasticity that challenges any intuitive notion of how living systems function, let alone survive and thrive in hostile environments. Biology is, in a word, messy.

Complexity seems to be the *mot du jour* in biology right now and the field arguably stands at a crossroads from which significant future progress will largely depend upon the ability of its researchers to at least tame the monster of complexity, if not to master it. Driven by the rapid



development of the technology available to scientists in the laboratory, the quantity, scope and resolution of biological data continues to grow at an accelerating pace. Invaluable though this burgeoning new wealth of data may be, it often poses more questions than it really answers until such time as it can be assimilated as real knowledge. The expectations for the Human Genome Project, for example, were as huge as the tsunami of data that it generated, spawning a whole new genomics industry within the life sciences sector and even moving some leaders in the field to make predictions about cures for cancer within a few years. Looking back over almost a decade since the first working draft of the human genome was completed[1] – while there have certainly been some medical benefits from this work, I think it is fair to say that the impact has not been on anything like the scale that was initially anticipated, largely as a result of having underestimated how difficult it is to translate such a large and complex body of data into real knowledge. Even today, our understanding of the human genome remains far from complete and the research to fill the gaps in our knowledge continues apace. The, as yet unfulfilled, promise of genomics is also reflected in the fact that many of the biotechnology companies that were founded with the aim of commercialising its medical applications have disappeared almost as rapidly as they arose. This is not to say that the Human Genome Project was in any way a failure – quite the opposite. In its stated goals to map out the human genome it succeeded admirably, even ahead of schedule, and the data that it generated will only become more valuable with time as we become better able to understand what it is telling us. The crucial lesson for biology, however, is that as our capacity to make scientific observations and measurements grows, the need to deal with the complexity of the studied systems becomes more, not less, of an issue, requiring the concomitant development of the means by which to synthesise knowledge from the data. Real knowledge is much more than just data – it does not come solely from our ability to make measurements, but rather from the intellectual frameworks that we create to organise the data and to reason with them.

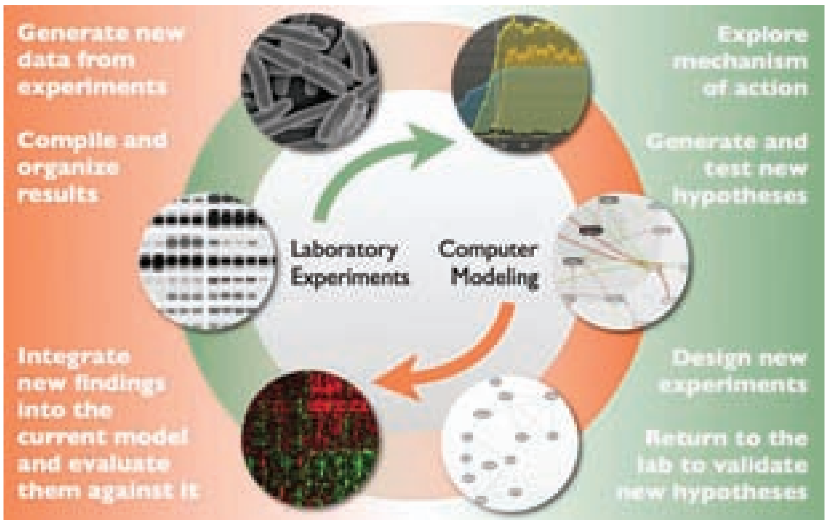
What are these intellectual frameworks? Modelling is one of the most powerful and widely used forms of intellectual framework in science and engineering. There is a tendency to think of models only in terms of explicit modelling – the creation of tangible representations of real world objects such as a clay model of a car in a wind tunnel or a simulation of a chemical reaction on a computer for example. In actuality however, scientists of all persuasions are (and always have been) modellers, whether or not they recognise this fact or would actually apply the label to themselves. All scientific concepts are essentially implicit models since they are a description of things and not the things themselves. The advancement of science has been largely founded upon the relentless testing and improvement of these models, and their rejection in the case where they fail as consistent descriptions of the world that we observe through experimentation. As in other fields, implicit models are in fact already prevalent in biology and are applied in the daily research of even the most empirical of biologists. Every experimental biologist who constructs a plasmid or designs a DNA primer for a PCR reaction, is working from an implicit model of the gene as a dimer of self-describing polymers defined by a four-letter alphabet of paired, complementary monomers. Interestingly, the double-helical structure of DNA[2] published by Watson and Crick in 1953, remained essentially just a hypothetical model until it could actually be observed by x-ray crystallography in the 1970s, yet it earned them a Nobel prize a full decade earlier as a result of the profound and wide-ranging synthesis and integration of knowledge that it brought to the field.

It is no accident that the majority of the successes that explicit modelling approaches have enjoyed in biology tend to be confined to a rather limited set of circumstances in which already established modelling methodologies are applicable. One example is at the molecular level where the quantitative methods of physics and chemistry can be successfully applied to objects of relatively low complexity by comparison with even a single living cell. Modelling approaches can also be successfully applied to biological systems that exhibit behaviour that can be captured by the language of classical mathematics, as for example in the cyclic population dynamics of the interaction of a predatory species with its prey, expressed in the Lotka-Volterra[3] equations. Unfortunately for the modern biologist, many if not most of the big questions in biology today deal with large, complex systems that do not lend themselves readily to the kind of modelling approaches just described. How do cells make decisions based upon the information processed in cell signalling networks? How does phenotype arise? How do co-expressing networks of genes affect one another? These are the kinds of questions for which the considerable expenditure of time, effort and resources to collect the relevant data typically stands in stark contrast to the relative paucity of models with which to organise and understand these data.

One response to the challenges posed by biological complexity in both academia and the healthcare industry has been the pursuit of more correlative approaches to biology in which an attempt is made to sidestep the complexity issue altogether by focusing (at least initially) on the data alone. If the data are measured carefully enough and can be weighted and scaled meaningfully with respect to one another, parametric divergences that can be detected between similar biological systems under differing conditions may reveal important clues about the underlying biology as well as identify the critical components in the system. This is very much a discovery process that aims in effect to pinpoint the needle in the haystack from the outside, without all the mess and fuss of having to get in and pull the haystack apart. These phenomenological approaches have been in widespread use for some time in both academia and industry in the form of genomic and proteomic profiling, biomarker discovery and drug target identification, and have also given rise to a plethora of high-throughput screening methodologies.

While these correlative approaches have enjoyed some successes, they do tend to compound the central problem alluded to earlier of generating data without knowledge. Moreover, their limitations are now starting to become apparent. In recent years for example, steadily intensifying investment and activity in these areas by drug companies has seen the approval rate of new drugs continue to fall as levels of R&D investment soar[4]. The field of biomarkers has also seen a similar stagnation, despite years of significant investment in correlative approaches. Cancer biomarkers are a prime example of this stagnation. Since we don't yet have a good handle on the subtle chains of cause and effect that divert a cell down the path towards neoplasia, we are forced to wait until there are obvious alarm bells ringing, signalling that something has already gone horribly wrong. The ovarian cancer marker CA125 for example, only achieves any significant prognostic accuracy once the cancer has already progressed beyond the point at which therapeutic intervention would have had a good chance of being effective[5]. To use an analogy from the behavioural sciences, broken glass and blood on the streets are the 'markers' of a riot already in progress but what you really need for successful intervention are the early signs of unrest in the crowd before any real damage is done. The lack of new approaches has also created a situation in which many of the biomarkers in current use are years or even decades old and most of them have not been substantially improved upon since their discovery. Much more useful than the current PSA test for prostate cancer[6], for example, would be a test with which a physician could confidently triage prostate cancer patients into groups of those who would probably require medical intervention and those whose disease is relatively quiescent and who would likely die of old age before their prostate cancer ever became a real health problem.

Given the general lack of useful mechanistic models or suitable intellectual frameworks for managing biological complexity, the tendency to fall back on phenomenology is easy to understand. Technology in the laboratory continues to advance, and the temptation to simply measure more data to try to get to where you need to be grows ever stronger as the barriers to doing so get lower and lower. We should also not underestimate the bias in the private sector towards activities that generate data, since they are far more quantifiable in terms of milestones and deliverables than the kind of research needed to translate that data into knowledge; this also makes these data-generating activities much more amenable to the kinds of R&D workflow management models that are common in many large biotechnology and pharmaceutical organisations. In effect what we have witnessed in biology over the last decade or so is a secular movement away from approaches that deal with underlying causation, in favour of approaches that emphasise correlation. However, true to the famous universal law that there's no such thing as a free lunch, the price to be paid for avoiding biological complexity in this way is a significant sacrifice with respect to knowledge about mechanism of action in the system being studied. Any disquieting


Generate new data from experiments
Compile and organize results
Integrate new findings into the current model and evaluate them against it
Laboratory Experiments
Computer Modeling
Explore mechanism of action
Generate and test new hypotheses
Design new experiments
Return to the lab to validate new hypotheses

feeling in the healthcare sector that it is probably a waste of time and money to simply invest more heavily in current approaches[7] is perhaps the result of an uneasy acknowledgement that much of the low-hanging fruit has already been picked and that any significant future progress will depend upon a return to more mechanistic approaches to disease and medicine.

Having arrived at such an intellectual impasse, there has arguably never been a more opportune time for biologists to incorporate modelling into their research. Biological modelling has to date tended to be almost exclusively the realm of theoretical biology, but as platforms for generating and testing hypotheses, models can also be an invaluable adjunct to experimental work. This is already well recognised in other fields where modelling is more established in the mainstream of research. Astronomers for example, use computer-based gravitational models to accurately position telescopes for observations of celestial objects, and civil engineers routinely test the load-bearing capacity of new structures using computers, before and during their construction, comparing the simulation results with experimental data to ensure that the limits of the structure's design are well understood.

One misconception that is common among scientists who are relatively new to modelling is that models need to be complete to be useful. Many (arguably all) of the models that are currently accepted by the scientific community are incomplete to some degree or other, but even an incomplete model will often have great value as the best description that we have to date of the phenomena that it describes. Scientists have also learned to accept the incomplete and transient nature of such models since it is recognised that they provide a foundation upon which more accurate or even radically new (and hopefully better) models can be arrived at through the diligent application of the scientific method. Physicists understand this paradigm well. For example, the discrepancies that astronomers observe trying to accurately map the positions and motions of certain binary star systems to the standard Newtonian gravitational model, can actually provide the means to discover new planets orbiting distant stars. The divergence of the astronomical observations from the gravitational model can be used to predict not only the amount of mass that is unaccounted for (in this case, a missing planet), but also its position in the sky. Models therefore, can clearly have predictive value, even when they diverge from experimental observations and appear to be 'wrong'.

For models to be successful in fulfilling their role as intellectual frameworks for the synthesis of new knowledge, it is essential that the chosen modelling system be transparent and flexible. If these requirements are satisfied, a partial model can still be of great value since the missing pieces of the model are a fertile breeding ground for new hypotheses and the model itself can even provide a framework for testing them. What is meant by transparent and flexible in this case? Transparency here refers to the ease with which the model can be read and understood by the modeller (or a collaborator). Flexibility is a measure of how easily the model can be modified, for example when new knowledge becomes available or existing knowledge turns out to be wrong. A model that is hard to read and understand is also difficult to modify and, very importantly in this age of interconnectedness, difficult to share with others. The importance of this last point cannot be overstated since one of the most often ignored and underestimated benefits of models is their utility as vehicles for collaboration and communication. A model that must be completely rewritten in order to add or remove a single component is not flexible and once built, there exists a significant barrier to modifying it that discourages experimentation. Flexibility also implies that the models should be capable of being built in a simple, stepwise fashion, based only upon what is known – free of any significant constraints on scope and resolution or any *a priori* hypotheses about how the assembled system will behave. The co-existence of these last two properties is essential for flexibility in a modelling system, since decisions about which components of a model to omit in order to keep it within workable limits (eg for the purposes of memory, storage or execution), must of necessity be predicated upon some pre-existing notions of how the assembled system will behave and which components will be critical for that behaviour. It is interesting to note that biological models based upon classical mathematical approaches generally fall far short of these ideals with respect to both transparency and flexibility.

A cornerstone of any modelling approach should also be the recognition that biological systems are dynamic entities – in fact 'biology' is essentially the term that we apply to the complex, dynamic behaviour that results from the combinatorial expression of their myriad components. For this reason, models that can truly capture the 'biology' of these complex systems are also going to need to be dynamic representations. This entails a big step beyond the computer-generated pathway maps that biologists working in cell signalling are already starting to use. These pathway maps are essentially static models in which the elements of causality and time are absent. They are a useful first step, but just like a street map, a pathway map does not provide any information about the actual flow of traffic. An ideal modelling platform would offer a 'Play' button on such maps, allowing the biologist to set the system in motion and explore its dynamic properties.

Notwithstanding these requirements for transparency, flexibility and dynamics, the biology community in large part are unlikely to adopt modelling approaches that require them to become either mathematicians or computer scientists, irrespective of how much math and/or computer science there might be 'under the hood'. How fewer cars would there be on our roads if all drivers were required to be mechanics? Ideally biologists will be able to couch their questions to these new frameworks in an idiom that approximates the descriptions of biological systems that they are used to working with – and in a similar fashion, receive results from them whose biological interpretation is clear.

Finally, let us not forget that thanks to the internet, we live in an era of connectivity that offers hitherto unimaginable possibilities for communication and collaboration. The monster of biological complexity is in all likelihood, too huge to ever be tamed by any single research group or laboratory working in isolation and it is for this reason that collaboration will be key. With knowledge and data distributed widely throughout the global scientific community, a constellation of tiny pieces of a colossal puzzle resides in the hands of many individual researchers who now have the possibility to connect and to work together as never before, and to assemble a richer and more complete picture of the machinery of life than we have ever seen. Potentially huge opportunities might therefore be squandered if future modelling platforms are not purposefully designed around the connectivity and semantic web capabilities that the internet offers. It is interesting to reflect on a potential future in which the monster of biological complexity, the emergent property of a multitude of tiny interactions within living cells, is eventually tamed by an analogous emergent property of the semantic web – the concerted scientific efforts of an interconnected, global research community.

Acknowledgement

With sincere thanks to Ty Thomson and Paul Edwards for proofreading and helpful discussion, and to Jean Krivine for the 'missing planet' analogy. DDW

A structural biologist by training, ***Gordon Webster*** *spent the decade after graduate school at academic research institutes in four countries, during which time he studied key components of cell regulation at the molecular level, taught extensively and ran his own research group. Since moving to the commercial sector, Gordon has used computers to design biotherapeutics for clinical use in oncology, and has worked on the development of advanced algorithms and applications in systems biology.*

References

1 International Human Genome Sequencing Consortium (2001) Nature 409: 860–921.
2 Watson, JD, Crick, FH (1953). Nature 171 (4356): 737-8.
3 Volterra, V (1931). in Animal Ecology (McGraw-Hill).
4 Belsey, MJ (2007). Nature Reviews Drug Discovery 6, 265-266.
5 Jacobs, IJ et al (1999). The Lancet, 353, 1207-1210.
6 Schröder, FH (2009). Recent Results Cancer Res. 181, 173-82.
7 Cuatrecasas, P (2006). J. Clin. Invest. 116, 2837-2842.